DOCUMENT RESUME

ED 144 117 CS 501 809

AUTHOR Farace, Richard V.; Mabee, Timothy
TITLE Communication Network Analysis Methods.
PUB DATE Apr 77
NOTE 33p.; Paper presented at the Multivariate Analysis in Communication Research Conference (Asilomar, California, April 1977) ; Best copy available

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Cluster Grouping; *Computer Programs; *Factor Analysis; Group Structure; Intercommunication; *Multidimensional Scaling; *Networks; *Speech Communication; *Statistical Analysis

ABSTRACT

This paper reviews a variety of analytic procedures that can be applied to network data, discussing the assumptions and usefulness of each procedure when applied to the complexity of human communication. Special attention is paid to the network properties measured or implied by each procedure. Factor analysis and multidimensional scaling are among the major approaches discussed; specific computer programs that are applicable to network data are also considered. (AA)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED144117

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Communication Network Analysis Methods

BEST COPY AVAILABLE

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Richard V. Farace
Timothy Mabee

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND THE ERIC SYSTEM CONTRACTORS."

Richard V. Farace
Timothy Mabee

Department of Communication
Michigan State University

February, 1977

First draft of a paper for the Multivariate Analysis Conference, Asilomar, California, April, 1977. For Comments Only.

CSSA1 889

The concept of a communication network has multiple origins in the literature on social networks. In sociometry, Moreno (1934) first interpreted patterns of social preferences between individuals as representing a network of human relations. Bavelas (1950) and Bales (1950) manipulated and measured respectively, the network of communication channels linking the members of small groups in social psychological laboratory settings.

Nearly 40 years later, Barnes (1972) described the conceptual and pragmatic importance of networks in the field of anthropology, and Bott (1971) used anthropological methods to trace and compare the different interpersonal networks of husband and wives. Recently, research in organizational communication has investigated the relationship of communication network properties to such variables as morale and satisfaction (Shaw, 1964, Wigand, 1974a, Danowski and Farace, 1974).

Interest in the concept of social networks, of which communication serves as one subset of social relations, implies a recognition of the complexity of social phenomena (cf. Monge and Day, 1976); "network" is essentially a multivariate concept, requiring multivariate analytic techniques to reduce what are often large collections of data to a small number of network properties. Festinger, Schachter, and Back observe: "Without an adequate representational technique of handling such data the analysis of the exact patterns of interconnections among members of a group is virtually impossible unless the group is very small." (1950, p. 132)

Monge and Day (1976) distinguish between two kinds of multivariate approaches to complex data. The first, which they call multivariate dependence models, divides the variables into independent and dependent sets, and uses the former to predict the latter. They call the second approach multivariate

interdependence models, or data reduction models. Here, no distinction is made among the variables; instead, the goal of the analysis is "to identify underlying simple structures which can parsimoniously represent the complexity in the data." (pp. 208-209)

Similarly, Krippendorff (1972) calls such analytic device "explicit simplification procedures," and points out that these procedures are comparable to the processes of verbal-conceptual abstraction: data of great complexity are reduced to a few simple concepts of practical utility.

Network analysis belongs in this second category of multivariate analysis. Its value lies in (1) its replicability, a property not shared by intuitive conceptual processes, and in (2) the utility of the resulting simplification as indicators or predictors of other social phenomena. Thus, the value of network analysis in understanding complex social phenomena depends not only on the rigor of the analytic procedure, but also on the uses found for the simplified representation of social structure produced by the analysis.

This paper reviews a variety of analytic procedures which can be applied to network data, and attempts to evaluate their usefulness for illustrating the complexity of human communication affairs. Special attention will be paid to the network properties measured or implied by each procedure, because these properties must ultimately take their place in a theory of social structure if the network concept is to evolve as fruitfully as its intuitive appeal suggests it could.

BASIC CONCEPTS

Some general definitions will be helpful for comparing different multivariate analytic techniques that apply to networks. First, the entities composing the network will be called nodes. In many applications, the nodes of the network are people, but studies have also been done of the networks of relationships among groups (Danowski and Farace, 1974), organizations

(Wigand, 1974b; Wigand and Larkin, 1975), among academic publications (Small, 1974), and among generalized role categories (Burt and Linn, 1977). Many other interesting applications are conceivable.

The measured relationships between individual nodes will be called links. Links are usually specified between pairs of nodes. If the relationship represented by a link is considered to be inherently mutual or symmetric, like a correlation, there is one link for every pair of nodes in the network. If the relationship is asymmetric, like a sociometric choice, there are two links for every pair of nodes. Hence, if a network has N nodes, it has $N(N - 1)/2$ symmetric pairwise links or $N(N - 1)$ asymmetric links.

Links have been measured at each of Stevens' (1951) four levels of measurement: nominal (the relationship is present or absent), ordinal (rank of the frequency or importance of the relationship for the initiating or receiving node), interval and ratio (strength or amount of the relationship). Occasionally, non-linear measures are also used, such as correlations, which form polar scales.

The values of links may be measured directly or derived from other measurements. When the nodes are individuals, the most common direct measurement technique is to obtain reports from the individuals within the boundaries of the intended network. Bott (1971), for example, used face-to-face interviews with family members to gather information on the contacts of family members with others. Jacobson and Seashore (1951) provided respondents with a list of potential network members, and asked each person on the list to evaluate the frequency and importance of their contact with every other person. Conrath (1972) asked members of an organization to keep diaries of their communication contacts with other members. Unobtrusive measures (see e.g., Martin and Ackoff, 1961) and mechanical means (Killworth and Bernard, 1976) have also been used to assess directly the links between network nodes.

Link measures derived from other data include correlations or other association measures applied to cross-sectional data on the traits or activities of nodes. For example, Phillips and Conviser (1972) treated nodes (persons) as variables and assessed the strength of their relationships based on their co-participation in fourteen social activities. Similarly, Small and Griffith (1974) measured the relationships between pairs of scientific publications as a function of the number of times each pair was cited together in another publication. Burt and Linn (1977) analyzed the content of newspaper articles for the co-occurrence of "actors," assigning each co-occurrence a weight based on the number of column inches devoted to the relationship, and adding the weighted co-occurrences to measure link strength.

It is often convenient and useful to distinguish several types of relationships between the nodes of a network. In Jacobson and Seashore's questionnaire, for instance, respondents separately reported their contacts on each of several topics, as well as specifying the medium of communication. Thus, each pair of nodes has not one, but potentially several distinct types of links between it.

Networks and analytic procedures which maintain distinctions between types of links are called polythetic, while those which use only a single type of link are called monothetic networks. Procedures which combine links of different types into one single type, by simple or weighted averaging, must be considered monothetic. No method of handling multiple types of links in a single analysis is now known to the authors; the development of a truly polythetic network procedure would be a major contribution to the methodology of network analysis.

A convenient way of arraying the pairwise links of a network is in the form of a square matrix, S, with each element s_{ij} corresponding to the link from the ith node to the jth node. The values of the diagonal elements s_{ij}, $i = j$, may be either one or zero or some other value, depending on the

requirements of a particular analytic procedure. If the relationship represented by the links is symmetric, and ones (1's) are used in the diagonal, the resulting matrix is also symmetric. The matrix representation of links is attributed to Forsyth and Katz (1946). It is isomorphic with the graph representation used by Moreno. Table 1 and Figure 1 show equivalent matrix and graphic representations of a network of three nodes using ratio-scaled asymmetric links.

TABLE 1

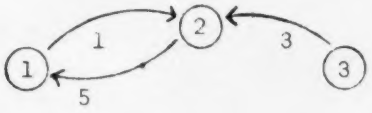
Hypothetical link strength matrix

Strength = frequency of communication per week

		Chosen Nodes		
		1	2	3
Respondent Nodes	1	X	1	0
	2	5	X	0
	3	0	3	X

FIGURE 1

Hypothetical sociogram corresponding to Table 1



NETWORK PROPERTIES

The common goal of all network analytic procedures is to decipher the underlying social structure implied by a particular link matrix. This general goal can be reached (or at least approached) by many different routes. Most procedures for network analysis are primarily directed towards decomposing the

network into groupings or cliques of highly associated nodes.* In using these procedures, the investigator characterizes network structure by the number and size of cliques present, and qualifies nodes in terms of their participation in cliques.

The criteria for detecting cliques are not well agreed upon, and tend to vary considerably across procedures for network analysis. Most analytic methods include, however, at least implicitly, some measure of the total amount of cohesion or structure in a subset of nodes as one criterion for the presence of a clique. When these measures are reported along with the clique membership results, the analysis becomes more useful, because it is possible not only to compare cliques within a network, and to compare entire networks quantitatively with regard to this fundamental property of amount of structure, but also to predict non-network variables, such as homogeneity of beliefs and attitudes (e.g., Danowski and Farace, 1974) from measures of network structure. Various approaches to defining indices for the amount of network structure in terms of a link matrix are discussed below.

Regardless of whether a particular investigation or procedure emphasizes the detection of cliques or the quantification of structure, or employs a combination of these two methods for representing social structure, the properties revealed by network analysis procedures can be broken into three broad categories:

(1) Properties of the network as a whole, including total amount of structure and various indications of the homogeneity of structure throughout the network. The latter property includes information about the number and

---

*See Holland and Leinhardt, 1975, and the discussion of Coleman's procedure, below, for exceptions.

size of detectable cliques, as well as the distribution of clique and node properties.

(2) Properties of sub-networks such as groups or cliques, including the amount of structure binding the members of each clique, and the number of members in each clique, and indicators of the relationship of each clique with other parts of the network.

(3) Properties of the sub-networks associated with each individual node, including the role of the node in the network (e.g., whether or not the node is a clique member), and the amount of structure binding each node to others in the network.

Since clique detection methods are in general based upon some index of structure, we will assume for the moment that cliques have been found, and proceed to consider in some detail several alternative measures of their structure. Such discussions apply equally well to the structure of the entire network, but require some modification when individual nodes are the focus.

Clique Structure Measures

One of the most satisfactory indices of clique or network structure, from a statistical point of view, is Coleman's (1964) measure of "hierarchization" in a nominal level measure of links displayed in an asymmetric matrix. The observed distribution of links is compared with its expected distribution if the links were to occur randomly. Coleman suggests, however, that measures which take into account only the distribution of links may leave out important aspects of the network which "may be of most importance for theoretical concerns." He recommends an alternative, based on work by Luce and Perry (1949), which assesses the "connectedness" of the network, or the ratio of the number of non-zero links to the total number of possible links in the network.

The measure can then extend to indirect, multi-step links; if a link matrix containing only ones and zeros is multiplied by itself, the result is a matrix containing the number of two-step paths from the i*th* node to the j*th* node, for all nodes. The ratio of the sum of these numbers to the number of possible two-step paths gives the two-step connectedness of the network. Coleman suggests continuing the multiplication operation until no new paths are added by the next multiplication. A summary measure of the connectedness density of the network is then the proportion of ultimately connected nodes, with respect to the total number of connections possible.

When metric information is available in the link matrix, other measures of clique or network structure are available. Of course, it is also possible to ignore the metric information by establishing a cutoff value, thus transforming the data into nominal terms, as above. A metric matrix, however, is naturally conceivable in spatial terms (Woelfel and Danes, 1977), suggesting *volume* as a measure of the structure in any subset of nodes. In a spatial representation, strong links are specified by *small* distances in the matrix, so volume is an inverse measure of structure, with respect to connectedness.

Comparison across networks requires the use of a common or standardized unit of measurement for the links. Volume is not, however, a standardized measure like connectedness. Volume depends not only on the unit used for measuring distances between nodes, but also on the number of nodes. For comparisons across networks and groups of different size, a derivative measure, *density*, is preferable. Density is defined as the number of nodes per unit volume.

Measures of Node Properties

The simplest way of characterizing the network properties of individual nodes is to take the average number or strength of links to and from each node. These measures, like the hierarchization measure discussed above, then,

appear to leave out key aspects of the sub-networks associated with each node. One alternative, which corresponds closely with the connectedness measure for cliques and networks, was proposed and implemented by Richards (1974c). Called "integrativeness," it measures the one-step connectedness density of those specific nodes which are directly linked with the starting node. The integrativeness of a node has a value of zero when none of the nodes which are linked with the starting node are linked with each other, and has a value of one when the nodes connected with the starting node are completely interlinked at the one-step level.

The metric equivalent of integrativeness is the density of nodes within a fixed radius of the target node.

The detection of cliques in a network makes possible a further characterization of individual nodes in terms of their roles in the clique structure of the network. For most applications of network results, it is desirable to arrive at an unambiguous assignment of the clique membership role of each node. It also seems advisable to allow for nodes to be members of no clique, at least when it is possible for a node to have no links. Finally, special roles may be assigned to nodes which have many links to more than one clique, thus serving as bridges and liaisons between cliques (see Richards, 1974a; Farace, Monge and Russell, 1977).

Finer investigation of the role of individual nodes within cliques has been pursued. MacKenzie (1966) has argued for the importance of a measure of the centrality of individuals in a small group, defining his index in terms of the distribution of links with other individuals. A similar measure in spatially represented networks is the distance of each node from the center of the clique, defined as the vector addition of distances to all other nodes in the clique.

CLIQUE DETECTION PROCEDURES

The general lack of agreement about the criteria for the detection of cliques in networks has a legitimate basis in the variety of existing conceptions and theoretic representations of social structure. Regrettably, the underlying assumptions are often left implicit, and must be inferred from the measurement and data reduction operations performed. While it is beyond the scope of this paper to perform the inference and make the connections with sociological and social psychological theory, we will attempt to bring into focus, albeit briefly, the measurement requirements and salient features of the operations performed in some of the major clique-detection procedures. In various applications, different conceptions and criteria may be appropriate; for the purpose of this review, the important factors are (1) whether the criteria are well-defined, and (2) whether they lead to interesting and potentially useful interpretations of cliques in terms of some research question.

Major differences in clique criteria exist between the nominal procedures, which characterize the sociometric research tradition (for instance, a node usually may be a member of only one clique), and spatial procedures (where multiple clique membership may be allowed). Within these two broad categories of procedures, differences in the meaning of clique arise because of variations in detection and clustering methods, and because of different methods for measuring links. Because of the many ways in which procedures can vary, direct comparisons and judgments of relative merit are difficult, but contrasts will be drawn whenever they seem appropriate.

In addition to detecting cliques and providing measures of structural properties for the whole network, for cliques, and for individual nodes, a procedure for network analysis is particularly useful in a practical sense if it can accomodate data with many nodes and produce results at low cost.

In describing specific procedures, we will try to indicate typical network size limitations for available computer implementations. Processing costs vary widely with computer installations and, for some procedures, with network properties as well, so it is not feasible to give cost estimates.

Nominal Clique-Detection Procedures

The simplest link-strength matrix is a binary matrix, containing only one or zero in each cell to indicate the presence or absence of a relationship between the nodes. Initial work on discovering cliques of highly interconnected nodes in such data was carried out by Luce and Perry (1949) and Luce (1950), and is described and applied to social networks by Festinger, Schachter, and Back (1950). Luce and Perry provide the following definition of a clique:

> A subset of the group forms a clique provided that it consists of three or more members each in the symmetric relation to each other member of the subset, and provided further that there can be found no element outside the subset that is in the symmetric relation to each of the elements of the subset. (1949, p. 97)

Nodes may be members of more than one clique.

Luce and Perry show that it is possible to find out which nodes are members of cliques as follows: First, make the link strength matrix symmetric by deleting unreciprocated links. Second, raise the symmetrized strength matrix to the third power (its cube). Third, examine the main diagonal of the cubed matrix. Nodes corresponding to non-zero diagonal elements are members of cliques.

This procedure does not allow the investigator to determine precisely how many cliques there are, nor which nodes are members of which cliques, but these goals can be accomplished in simple cases by returning to the symmetrized strength matrix, and visually examining the specific patterns of symmetric relationships between the nodes which are known to be members of cliques. A more serious difficulty with this approach, mentioned by the

authors, is that more complex structures (which by intuition one might like to consider as cliques--for instance, sets of nodes which meet the criteria cited above except for a few unreciprocated or missing links) cannot be considered. Luce (1950) addresses this problem in a later paper, basing a generalized definition of an n-clique which uses the concept of multi-step connectedness discussed in the preceding section of this paper. A set of three or more nodes is considered an n-clique if paths of n or fewer steps connect all members of the set. A matrix multiplication procedure is used to detect nodes which are members of n-cliques. The problem of finding specific clique membership remains, however.

Harary and Ross (1957) address the problem of specific clique membership by showing that in networks with fewer than four cliques, at least one node in each clique must be a member of only that one clique. They give criteria for finding these "unicliqual" nodes, and point out that all nodes connected with each such node at n or fewer steps are members of a distinct clique. They are able to generalize their results by induction to networks having an indefinite number of cliques.

A different, but related, conception of connected clique is proposed by Weiss and Jacobson (1955). Instead of weakening the complete, single-step connectedness criterion of Luce and Perry by formally allowing additional steps to complete the connections between clique members, they simply permit cliques to be less than fully connected. Their detection procedure consists of reordering the rows and columns of a symmetrized link matrix so that nodes who have contacts with each other are adjacent in the matrix. After "liaisons" are removed, the remaining nodes often fall into visibly separate, dense blocks, which they define as mutually exclusive cliques.

They attempt to resolve less dense "grey" patterns by tentatively reclassifying key nodes as liaisons and reordering the nodes again, in an attempt to decompose the amorphous areas of the matrix into clearer, more dense aggregations.

In spite of the lack of specific, replicable criteria for both the reordering and the identification of cliques "by inspection" (both operations are carried out by hand), the Weiss and Jacobson method has interesting implications. It provides, for the first time, the possibility of assigning roles to the nodes, in terms of the clique structure of the network. Any node may be a member of one clique or may be an isolate, or may be a liaison by virtue of strong connections with more than one clique. Further, their procedures seem susceptible of rigorous formulation, though they themselves do not provide such a formulation. Richards' (1974b, 1975) NEGOPY program (see the last section of this paper) uses and formalizes many of Weiss and Jacobson's ideas.

Other methods, some using criteria for clique detection which are unrelated to the connectedness criterion emphasized here, may also be used with binary network data. However, most of these methods also permit the use of metric (ordinal, interval, or ratio-scaled) data, and hence will be dealt with in the next section. When binary data are used in such procedures, the data may be seen as possessing metric information with very low resolution.

Metric and Ordinal Clique-Detection Procedures

The forerunner of the spatial procedures is factor analysis. The link strength matrix for factor analysis of networks must be in the form of a correlation matrix. One possibility for the source of the matrix was mentioned in the section on link measurement: correlations are calculated from cross-sectional co-occurrence data, using the nodes as variables and traits or activities as cases. Glanzer and Glaser (1959) describe a slightly different approach, involving the computation of correlations between the row vectors of an existing asymmetric link matrix. This approach is a special case of the one suggested first, with "chooses node j" as an activity.

Note that the analysis of correlation matrices derived in either way implies that the investigator is interested in a network based on similarities between the nodes with respect to traits, activities, or previously defined links. In the case of activities, an argument may be made for interpreting the correlations as indicators of an interdependence relationship, but the argument is rather weak for the activity of making a sociometric choice, which, after all, is stimulated by the investigator, rather than by the transmission of influence in the social setting of interest.

Yet another way to arrive at a suitable matrix for factor analysis, and avoid the similarity connotation in favor of a connectedness approach, is to measure the links in such a way as to form a polar scale with values between minus one and one. If symmetric links are originally measured on a ratio

scale, one transformation* which might be used to achieve the desired polar scale is:

$$s'_{ij} = \begin{cases} \sqrt{\dfrac{s_{ij}}{\max\ [S]}}, & i \neq j \\ 1, & i = j \end{cases}$$

which is a polar scale between zero and one.

If the symmetric correlation or S' matrix is non-singular, it is orthogonally decomposed by factor analysis, i.e., transformed so that the input correlations are maintained (if principal components factor analysis is used), but the locations of the nodes are expressed in terms of an orthonormal basis (the N factors), rather than in terms of each other. Factors have been interpreted as cliques (see e.g., Killworth and Bernard, 1973), but the problem of how many factors to use must be solved first, since there are as many factors as nodes in the decomposed matrix. The scree test is one means for aiding this decision process.

Factor loadings give the extent of involvement of each node with each factor; an arbitrary minimum value must be chosen (Killworth and Bernard use 0.6) or else every node will be a "member" of every factor. Even with a high cutoff, unique assignment of nodes to cliques is not guaranteed.

Oblique factor rotations allow the assessment of relationships between cliques in the same metric as the input matrix. The strength of an inter-clique relation is the cosine of the angle between the factors corresponding to each clique.

Standard factor analysis programs will handle up to about 100 nodes.

---

*This transformation is not proposed as a practical expedient, but rather to clarify the relationship of factor analysis with other clique-detection methods which use a strength or distance matrix.

Factor analysis has been presented here primarily as an introduction to the concepts involved in multidimensional scaling (MDS). For metric MDS (MMDS), the links must be ratio scaled and (usually) symmetric. Since MMDS operates on a matrix of distances between nodes, rather than link strengths, a transformation of the S matrix is necessary to prepare the data. The transformation is of the form:

$$d_{ij} = \begin{cases} 1/s_{ij}, & i \neq j \text{ and } s_{ij} \neq 0 \\ 0, & i = j \\ M, & s_{ij} = 0,\ i \neq j, \end{cases}$$

where M is a constant, $M > 1/s_{ij}$ for all i and j.

MMDS, using algorithms very similar to those of factor analysis, operates on the input distance matrix to place each node in a multidimensional space. The distance between every pair of nodes in the space is the same as the input distance. This sometimes results in a space with "imaginary" dimensions, or equivalently the "warped" spaces of Riemann geometry and tensor calculus (see Woelfel and Danes, 1977). There is nothing "wrong" with imaginary dimensions, except that many scaling methods discard them as irrelevant or uninterpretable, thus distorting the distances between some pairs of nodes in the space. Since imaginary dimensions are quite likely in MMDS results for networks (depending somewhat on the value chosen for the constant, M), the use of a scaling technique which allows such peculiarities is recommended. One such technique is the GALILEO program. GALILEO will accept up to 45 nodes.

The interpretation, along with the disadvantages, of MMDS as a clique detection device are similar to those of factor analysis: each of up to N - 1 dimensions may be a clique, and each node may be associated with

each clique. The problem of determining how many dimensions to retain remains; see Barnett and Woelfel (1976) for a systematic discussion of some solutions, many of which are also applicable to factor analysis.

While MMDS in itself seems to have limited value as a clique detection device, it appears to have application in the mapping of the "fine structure" within cliques discovered by other means, and perhaps in describing the relationships between cliques. Further, it has the advantage of considerable methodological development in the area of graphically plotting changes in the distance matrix over time (see Woelfel and Danes, 1977), a crucial and, until recently, neglected aspect of network analysis.

There exist a number of non-metric or partially metric multidimensional scaling methods which may also be used for network analysis, with similar interpretations. Smallest space analysis, for instance, produces a spatial image of the network in which only the ranks of the input distances are respected (Lingoes, 1972). It should be noted that data in which the respondents rank their links with others, rather than making interval or ratio judgements, do not meet the input requirements of SSA, since it preserves overall ranks rather than within-respondent ranks. For data which are ordinal to begin with, non-metric scaling methods should be used. The interested reader is advised to consult the collection of MDS articles edited by Shepard, Romney, and Nerlove (1972).

Interpreting the dimensions of a space as cliques entails a considerable distortion of the usual sociometric meaning of a clique as a set of highly interconnected nodes. The nodes which lie on a line in space may still be very far apart (weakly linked). The distance matrix and its orthogonal decomposition contain far more than dimensional information, however: it might be useful to consider the spatial representation of a network

as a detailed map, with the nodes as scattered houses, and the village boundaries not yet drawn in. By this analogy, the main task in identifying cliques becomes one of finding clusters of closely-spaced nodes in the spatial "map" of the network.

Among the best-known algorithms for finding clusters in multidimensional space are Ward's (1963) method and two methods developed by Johnson (1967). All are agglomerative hierarchical methods; they add nodes to existing clusters in successive passes, so as to produce a tree structure. Eventually all nodes end up in a single cluster; the problem is where to stop. See Anderberg (1973) for a discussion of several solutions for this difficulty. When a stopping point is reached, the clusters are mutually exclusive, and some distant nodes may remain isolated.

Cluster analysis detects cliques which are highly structured in terms of the measures proposed in the section on network properties, since by definition clusters are relatively densely populated portions of the multidimensional space. All of the spatially-derived properties of networks, cliques, and nodes are consistent with the structural implications of these methods. In addition, it seems possible and interesting in this context to investigate in some detail the relationships between cliques, perhaps in terms of closest, farthest, and average distances between members of different cliques.

In view of the apparent applicability of spatial clustering methods, it is surprising that they have been little used for the analysis of network data, although recent exceptions do exist. (See Breiger, Boorman, and Arabie, 1974, and the description of the SOCK program, below.) Perhaps this is due to the fact that practical implementations of clustering methods are fairly new (Blashfeld, 1976) and because much of their development and

use has been in the fields of biology and cognitive psychology. A typical clustering program has a capacity of 100 nodes, though Small (1974) evidently has access to a program of much higher capacity.

Special Network Procedures

We turn now to a collection of procedures designed specifically for the analysis of sociometric and other linkage data. The first such procedure is Breiger's CONCOR, which the authors have extensively compared with some of the scaling and clustering procedures above (Breiger, Boorman, and Arabie, 1974).

CONCOR is essentially a metric hierarchical clustering algorithm. Unlike the examples of general-purpose clustering methods of the preceding section, it partitions the nodes into successively smaller groups, rather than assembling them into larger and larger ones. No specific criteria are given for stopping the partitioning.

A key step in the CONCOR procedure is the calculation of correlations between the rows (or columns) of the link strength matrix. This transforms the original linkage data into what Shepard (1972) calls "profile" data. It produces clusters based on the similarity of nodes, rather than their linkage, as in the factor analysis of similar correlation matrices.

Though the authors tie this feature of their algorithm to a theory of sociometric data (Lorrain and White, 1971; see also White, Boorman, and Breiger, 1976), in the present context CONCOR would seem to be more applicable to (not necessarily square) matrices of trait or activity data such as that discussed by Phillips and Conviser (1972). CONCOR will accept unsquare matrices, and should be very useful for discovering similarity clusters in such data. A recent analysis of the mathematical foundations of CONCOR shows that the program produces results which are formally similar to those of factor analysis (Schwartz, 1977).

CONCOR is solely a group-detection device, and offers no measures of network properties, though if the results can be related to a metric space, the volume and distance measures associated with conventional cluster analysis could be used. This step seems doubtful, however, since the similarity measure used (correlation) is non-linear. Two of the authors of the CONCOR description discuss structural measures in a separate context (Boorman and Arabie, 1972).

A particularly interesting feature of CONCOR is its ability to handle multiple kinds of association between the nodes, referred to above as polythetic data. However, the three-dimensional polythetic data array is immediately collapsed by the program into an unsquare two-dimensional matrix, which appears from the authors' description to have the same effect as taking an unweighted sum or average of the different types of links, a feature which is available to any of the network procedures discussed, merely by preprocessing the data. Thus CONCOR is not truly polythetic, in the sense of maintaining the distinction between types of links through the cluster-finding stage.

The node capacity of CONCOR is not known at present.

Killworth and Bernard's (1973) KBPAK is a collection of programs designed to process ordinal sociometric data of a particular type. (See Killworth and Bernard, 1974a and 1974b, for descriptions and evaluations of the data-gathering method.) The essential feature of the data is that it is ranked within informants; i.e., each respondent ranks his/her frequency of contact with all other nodes; the node with the highest frequency receives the lowest rank, so the data matrix is more an ordinal distance matrix, than a link strength matrix. After preprocessing to find the minimum distance between any pair of nodes, regardless of the number of

steps involved (i.e. if two nodes who do not have much contact are each linked closely with a third node, the sum of the "short" links with the third node, plus a small penalty for the extra step, replaces the "long" link between the first two nodes), correlations are computed between the rows of the matrix, and the resulting correlation matrix is factor analyzed. Nodes having a loading of 0.6 or greater on any factor retained (how many are retained is not clear) are treated as a "tentative subgroup." Nodes with a loading of 0.3 or more on two or more factors are candidates for a special subgroup-linking "intermediary" role. Because correlations are computed, all results must be interpreted in terms of similarity rather than linkage.

Two methodological criticisms can be made of the KBPAK procedures. First, though the program description does not specify, it is assumed that the ordinary product-moment correlation is used. This seems unjustified in light of the ordinal scaling of the data; perhaps the Spearman rank-order correlation should be substituted. Second, it is difficult to understand the minimum-distance procedure, since factoring is based on the similarity of node preferences; the preprocessing distorts those preferences in an unknown fashion. Further, since the distances are ranked within respondents, there is no common metric for comparing them across nodes, as required by the minimum-distance procedure.

A major point in favor of the KBPAK method, on the other hand, is its handling of some of the difficulties inherent in the factor analysis of networks. Ambiguity in assignment of nodes to cliques is greatly reduced, though not necessarily eliminated, by the high cut-off for factor loadings. More important is the creation of the special "intermediary" role, which allows for some consideration of the relations between cliques, even in the

absence of a specific measure of the strength of these relations. Also interesting is the authors' insistence on the tentativeness of cliques and role assignments, and the importance of supplementary anthropological description of the network (Bernard and Killworth, 1973). This position may be compared with the position taken earlier in this paper that network properties, if they are to be generally useful concepts, must be related to other variables.

The KBPAK series will accept up to about 150 nodes, depending on the memory capacity of the computer used. It has been designed for easy portability between computing sites.

Alba and Gutman (1971) have written a package, called SOCK, for the analysis of binary sociometric data. Based on Luce and Perry's (1949) and Coleman's (1964) concept of multi-step paths (see the discussion of network structure properties, above), the program constructs one of several optional pairwise distance measures from either asymmetric or forced-symmetry data. The resulting distance matrix is then clustered by one of Johnson's hierarchical clustering methods, or mapped into a multidimensional space by a version of Kruskal's MDSCAL. Thus SOCK offers a combination of several of the spatial procedures discussed above, together with linkage-derived preprocessing to achieve the necessary ratio scaled distances.

The key to understanding cliques or multidimensional map produced by SOCK is the derivation of the distance measures. The first option for computing distances from the nominal linkage matrix is a weighted sum of the number of paths between two modes, $d_{ij} = \sum_k a^k p_k$, where a is a constant chosen by the user (usually $0<a\leq 1$), and P_k is the number of paths of length k between the nodes. Note that the average of this measure over all nodes would be a measure of total structure, which is closely related to Coleman's

suggested network connectedness measure.

The second option is more complex, and will only be described in general here. In this option, d_{ij} is made inversely proportional to the degree of overlap of two circles of radius k around nodes i and j, where k is again the number of steps in a path between the two nodes. Notice that when k=1, the measure is a symmetric, inverted version of Richards' integrativeness measure.

Alba and Gutman do not mention any measures of network properties produced by their program, but several could easily be devised, based on information generated within the program. They give an example in which 1000 nodes are processed, but capacity depends on available core memory.

Richards' (1974a, 1974b, 1974c, 1975) NEGOPY program processes ratio scaled linkage measures in either asymmetric or forced-symmetry forms. The clique-finding procedures consists of two stages. The initial detection stage is a unidimensional clustering procedure which is recognizably, if distantly, related to the matrix manipulation methods of Luce (1950). This stage has several user-adjustable parameters, which serve to emphasize its heuristic origins. Cliques tentatively identified by the first stage are then subjected to a series of formal criterion tests. The formal criteria are as follows:

(1) Each node must have at least p percent of its total link strength within the clique (p must be greater than 50% for unambiguous membership role assignment).

(2) The clique must be completely connected, using paths of length q or less.

(3) Removal of "critical" nodes and links must not cause the clique to become disconnected at the chosen path length.

In addition to detecting cliques, the NEGOPY program provides a number of network property measures, including the one-step connectedness density of cliques, and the integrativeness of individual nodes, discussed above. Connectedness is not calculated for the network as a whole, perhaps because the large number of nodes the program is designed to handle might make such a calculation expensive. Node roles permitted include not only "clique member" and "isolate," but also "liaison" (non-member node with strong ties to two or more cliques), and "bridge" (clique member with strong ties to another clique).

NEGOPY is designed to accomodate up to 4095 nodes. It is not portable, and is currently available at only two Control Data Corporation (CDC) computer facilities.

PROBLEMS

The discussion so far has assumed that a social network is a thing unto itself, capable, perhaps, of being related to other social phenomena, but existing full-grown and unchanging in the social milieu. This assumption simplified the discussion of existing methods for network analysis, but it begs two questions. First, how do social networks arise, and second, how do they change in response to external stimuli?

The second question has a more straightforward response in terms of the concepts developed in this paper, so we will consider it first. It is quite conceivable, though few examples will be found in the sociometric literature, to compare the properties of several networks defined by the same boundaries, but separated in time. For the summary measures defined on the entire network or on individual nodes, the statistical problem of how to assess the changes is straightforward, capable of solution by standard time-series regression methods (see, for introductory treatments,

Nelson, 1973, and Hibbs, 1974, 1977), or by repeated measures analysis of variance (Finn, 1969; Poor, 1973).

Of special interest would be attempts to predict changes in the network properties from external events, such as changing economic conditions or management styles. The data required for such predictions would be massive; at least twenty points in time are necessary for most time-series analyses. The limited budgets and short-term nature of most social science research are major obstacles, but it seems obvious that networks must be studied in this fashion if the network concept is to attain theoretical status commensurate with its intuitive appeal.

At the level of the clique, the problem of change over time is much more difficult to deal with. How does one compare one configuration of cliques with another? If the changes are small, so that each clique at time one is readily identified with a similar clique at time two, a variety of change measures are conceivable, including Cohen's (1960) kappa statistic for classification agreement.

But suppose a clique decomposes over time into splinter groups? How, aside from observing that the decomposition has occured, can the change be measured? Further, it is not impossible that no such simple observation will fit the case--that the new configuration will bear no apparent relationship to the old. Part of the answer to this difficulty may eventually come in directions indicated by Woelfel and Danes (1977) on the comparison of multidimensional spaces over time. It is clear that most of the procedures for network analysis require much development in this regard, though it should be noted that Killworth and Bernard's KBPAK series of programs provides for the analysis of small changes in clique membership over a single time interval.

The question of how networks originate appears to be more of a conceptual question than an analytic problem at present. In collections of individuals, we may conceive that relatively stable network patterns arise as a consequence of a series of repeated communication events or social contacts. Some network investigations implicitly recognize this origin by asking respondents how frequently they talk to each other. A better example, though not a network study per se, will be found in Katz and Lazarsfeld's (1955) investigation of interpersonal mediation of mass media influence. The investigators make a serious attempt to refer reported contacts to specific times and places, so as to limit their inference to particular events. Their procedure in itself is a powerful check on the reliability of self-reported contacts, and serves as well to clarify the validity issues in self-reported network data. But only very recently, in the area of small group research, is the possibility being investigated that the sequence of events may also be important in determining the social relationship between individuals.

A third issue that needs to be raised concerns the monolithic emphasis in network analysis on pairwise relationships. It should be clear that when social relationships are measured between two nodes at a time, an assumption is being made that when three people meet, the relationship between each of the three pairs is the same as when each pair meets separately. Such an assumption seems very difficult to justify.

Finally, it should be observed that most network analysis procedures assume perfect measurement: all nodes within the network boundary are assumed to have been measured, and each measurement is assumed to be unbiased and perfectly reliable. This assumption is untenable for most data

sets. Since most of network analysis has no statistical basis,* the principal way of evaluating the effects of sampling and of random measurement error is by the Monte Carlo method, in which the results of many analyses with randomly disturbed data are compared to the "true" results. Some such work has been done on cluster algorithms by Blashfield (1976) which indicates that the effects of sampling, plus a random measurement error of up to 0.6 standard deviations, may be considerable. Johnson's two methods achieved median coefficients of agreement in classifying nodes of .06 and .42, while Ward's minimum-variance method achieved the highest agreement, a median of .77.

Blashfield's results are probably not typical of performance expectations of network procedures in the face of actual reporting errors encountered. For one thing, evidence exists that self-reports of social contacts are systematically biased, with respondents tending to over-report contacts with higher-status individuals, and to under-report contacts with individuals of lower status than themselves (Farace, Monge, and Russell, 1977). Killworth and Bernard (1976) have taken an interesting approach to the problem of self-report error. They have collected a pair of data sets for one group of individuals; one set was collected by conventional self-report methods, and the second set by mechanical means which are presumably free of error. The differences between the two sets are substantial; not surprisingly, a program for assessing network structures (Holland and Leinhardt's triad census, 1975) produces very different results from the two sets of data. The authors observe that the "true" data set produces no recognizable

---

*Two important exceptions have recently appeared. Granovetter (1976) shows how sample estimates of network connectedness can be made in very large networks, and Burt and Linn (1977) use multiple indicator models to assess link measurement validity and reliability.

structure, while the self-reported "cognitive" results are highly structured and interpretable.

REFERENCES

Anderberg, M.R. Cluster Analysis for Applications. New York: Academic
1973 Press.

Alba, R.D. & M.P. Gutman. SOCK--a sociometric analysis system. Mimeo.
1971

Bales, R.F. Interaction Process Analysis: A Method for the Study of
1950 Small Groups. Cambridge, Mass.: Addison-Wesley.

Bavelas, A. Communication patterns in task-oriented groups. In The
1950 Policy Sciences, Lerner and Lasswell (eds.).

Barnes, J. Social networks. Addison-Wesley Module #26, 1-29.
1972

Barnett, G.A. & J. Woelfel. On the dimensionality of psychological pro-
1976 cesses. Paper presented at the annual meeting of the Psychometric Society.

Bernard, H.R. & Killworth, P.D. On the social structure of an ocean-going
1973 research vessel and other important things. Social Science Research, 2:145-184.

Blashfield, R.K. Mixture model tests of cluster analysis: accuracy of
1976 four agglomerative hierarchical methods. Psychological Bulletin, 83: 377-388.

Boorman, S.A. & P. Arabie. Structural measures and the method of sorting.
1972 In Multidimensional Scaling, Shepard et al. (eds.).

Bott, E. Family and Social Network. (2nd edition) Glencoe, Ill.: The
1971 Free Press.

Breiger, R.L., S.A. Boorman, & P. Arabie. An algorithm for blocking rela-
1974 tional data, with applications to social network analysis and comparison with multidimensional scaling. Technical Report no. 244, psychology and education series. Stanford, CA: Stanford University Institute for Mathematical Studies in the Social Sciences.

Burt, R. & N. Linn. Network time series from archival records. In D. Heise
1977 (ed.), Sociological Methodology 1977. San Francisco: Jossey-Bass.

Coleman, J.S. Introduction to Mathematical Sociology. New York: Free
1964 Press.

Cohen, J. A coefficient of agreement for nominal scales. Educational
1960 and psychological measurement. 20: 37-46.

Danowski, J. & R.V. Farace. Communication network integration and group
1974 uniformity in a complex organization. Paper presented to the International Communication Association.

Farace, R.V., P. Monge, and H. Russell. Communicating and Organizing.
1977 Reading, Mass.: Addison-Wesley Press.

Festinger, L., S. Schachter, & K. Back. Social Pressures in Informal
1950 Groups. Stanford, CA.: Stanford University Press.

Finn, J.D. Multivariate analysis of repeated measures data. Multivariate
1969 Behavioral Research, 4: 391-413.

Forsyth, E. & L. Katz. A matrix approach to the analysis of sociometric
1946 data: a preliminary report. Sociometry, 9: 229-235.

Glanzer, M. & R. Glaser. Techniques for the study of group structure and
1959 behavior: I. Analysis of structure. Psychological Bulletin, 5: 317-332.

Granovetter, M. Network sampling: Some first steps. Am. J. Sociology,
1976 81: 1287-1303.

Harary, F. & I. Ross. A procedure for clique detection using the group
1957 matrix. Sociometry, 20: 205-215.

Hibbs, D.A. On analyzing the effects of policy interventions: Box-Jenkins
1977 and Box-Tiao versus structural equation models. In D. Heise (ed.). Sociological Methodology 1977. San Francisco: Jossey-Bass.

Hibbs, D.A. Problems of statistical estimation and causal inference in time-
1974 series regression models. In Sociological Methodology 1973-74, Costner (ed.). San Francisco: Jossey-Bass.

Holland, P. & S. Leinhardt. The statistical analysis of local structure in
1975 social networks. In Sociological Methodology 1975, Heise (ed.). San Francisco: Jossey-Bass.

Jacobson, E. & S.E. Seashore. Communication practices in complex organi-
1951 zations. Journal of Social Issues, 7: 28-40.

Johnson, S.C. Hierarchical clustering schemes. Psychometrika, 32: 241-254.
1967

Katz, E. & P.F. Lazarsfeld. Personal Influence. New York: The Free Press.
1955

Killworth, P.D. & H.R. Bernard. Catij: a new sociometric and its appli-
1974 cation to a prison living unit. Human Organization, 33:

Killworth, P.D. & H.R. Bernard. The catij technique: some descriptive
1974 tests of its adequacy. Paper delivered at the meetings of the American Sociological Association.

Killworth, P.D. & H.R. Bernard. The KBPAK--a collection of programs for
1973 the analysis of a social network. Office of Naval Research report no. BK-101-73.

Killworth, P.D. & H.R. Bernard. Informant accuracy in social network
1976 data. Human Organization, 35: 269-286.

Krippendorff, K. An algorithm for simplifying the representation of
1972 complex systems. Paper presented at the Internation Congress of
Cybernetics and Systems, Oxford, England.

Lingoes, J.C. A general survey of the Guttman-Lingoes Non-metric program
1972 series. In Multidimensional Scaling, Shepard, et al. (eds.).

Lorrain, F. & H.C. White. Structural equivalence of individuals
1971 in social networks. J. Mathematical Sociology, 1: 49-80.

Luce, D. Connectivity and generalized cliques in sociometric group
1950 structure. Psychometrika, 15: 169-190.

Luce, D. & A. Perry. A method of matrix analysis of group structure.
1949 Psychometrika, 14: 95-116.

Mackenzie, K.D. Structural centrality in communications networks.
1966 Psychometrika, 31:

Monge, P.R. & P.D. Day. Multivariate analysis in communication research.
1976 Human Communication Research, 2: 207-220.

Moreno, J.L. Who Shall Survive? A new approach to the problem of human
1934 interrelations. New York: Beacon.

Poor, D.D.S. ANOVA for repeated measures: Two approaches. Psychological
1973 Bulletin, 80: 204-209.

Nelson, C.R. Applied Time Series Analysis for Managerial Forecasting.
1973 San Francisco: Holden-Day.

Phillips, D.P. & R.H. Conviser. Measuring the structure and boundary
1972 properties of groups: some uses of information theory. Sociometry,
35: 235-254.

Richards, W.D. Network analysis in large complex systems: theoretical
1974a basis. Paper presented to the meetings of the International Communication Association.

Richards, W.D. Network analysis in large complex systems: techniques
1974b and methods. ICA.

Richards, W.D. Network analysis in etc.: metrics. ICA.
1974c

Richards, W.D. A manual for network analysis. A report of the Institute
1975 for Communication Research, Stanford University.

Schwartz, J.E. An examination of CONCOR and related methods for blocking
1977 sociometric data. In D. Heise (ed.) Sociological Methodology
1977. San Francisco: Jossey-Bass.